PewResearchCenter

Using timestamps for the evaluation of data quality

Gijs van Houten, *Research Manager, Eurofound*

Steve Schwarzer, *Research Methodologist, Pew Research Center*

Research questions

How are timestamps distributed?

To what extent do cases that were and were not flagged based on timestamps differ

- on other data quality indicators?
- in terms of the sample composition?
- in terms of the answers to substantive questions?

Survey design

- 17 countries from a cross-national survey

- Design:
  - CATI (10 countries): RDD, landlines and cell phones
  - CAPI (7 countries): Multi-stage stratified random sample, random walk

- Implementation
  - Conduct survey via coordinating vendors and local vendors
  - Involvement and oversight at every stage

- Social and political questions
  - CATI: ~20 minutes
  - CAPI: ~40 minutes

Quality control indicators

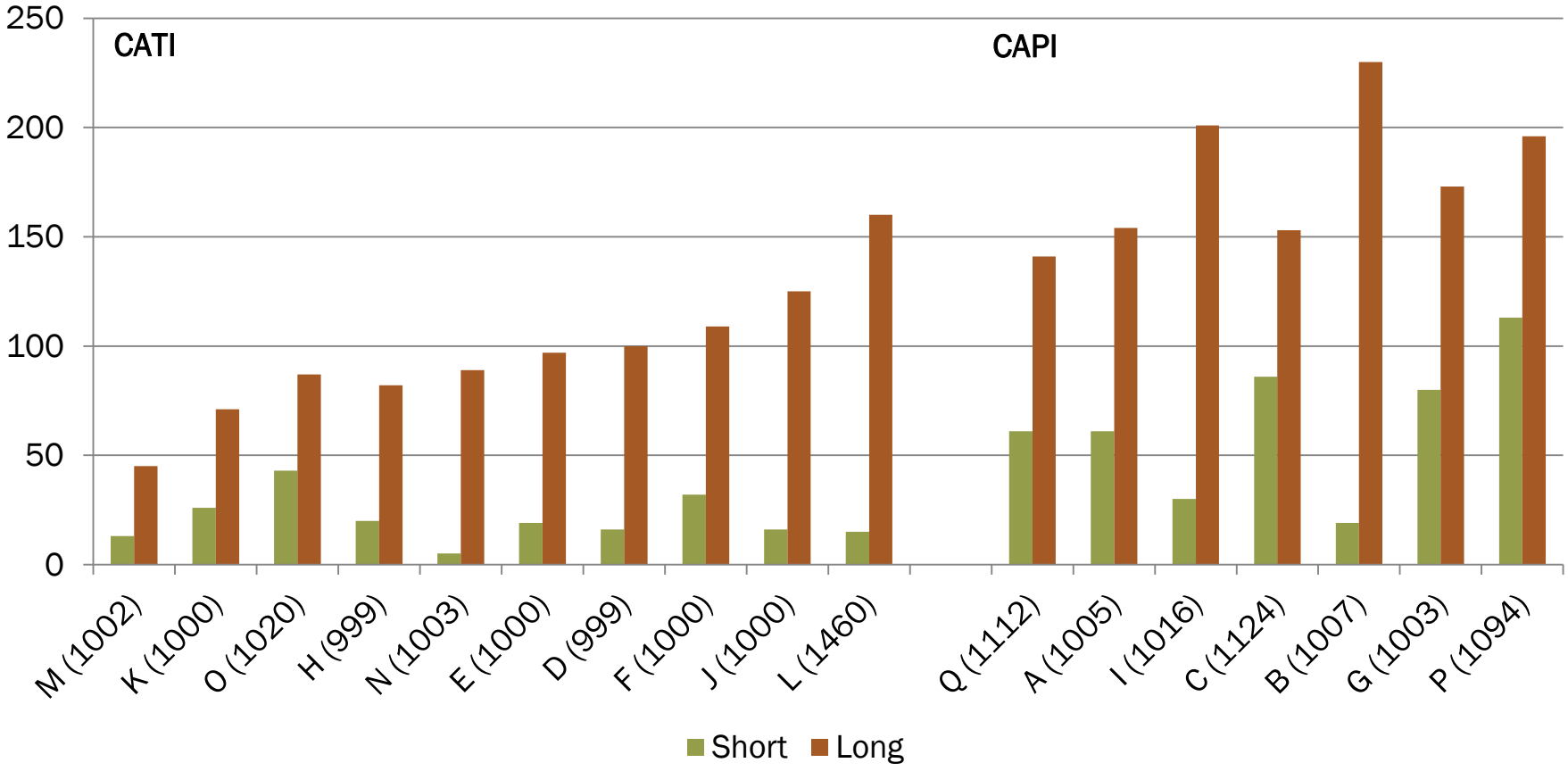
- **Paradata**
  - Timestamps
  - Interviewer workload
  - Overlapping interviews
  - Interviews in multiple PSUs (F2F only)
  - Interviews between 10 pm and 7 am
  - Interview duration
    - Short and long

- **Substantive data**
  - Item nonresponse
    - Low and high
  - Straight lining
  - High matches

Timestamps allowing timing of 6 blocks of items

		Maximum number of items						
Mode	Country	Block 1	Block 2	Block 3	Block 4	Block 5	Block 6	Total
CATI	N	9	13	6	12	14	23	77
	J	17	18	17	5	16	11	84
	H	18	13	20	15	12	11	89
	L	20	13	20	14	12	11	90
	D	18	13	20	15	12	12	90
	M	19	13	20	15	12	11	90
	K	19	13	20	15	12	12	91
	E	21	13	20	15	12	11	92
	O	17	13	23	14	8	18	93
	F	16	15	25	8	12	18	94
	Average	*17*	*14*	*19*	*13*	*12*	*14*	*89*
CAPI	I	19	13	20	16	31	11	109
	C	18	12	18	20	36	10	114
	P	19	12	18	18	37	10	114
	Q	18	12	21	18	35	10	114
	B	20	13	20	17	42	11	123
	A	20	15	34	16	40	12	137
	G	22	15	34	16	44	11	142
	Average	*19*	*13*	*24*	*17*	*38*	*11*	*122*

Items flagged for extreme timings



Associations with other QC indicators – short timestamps

		Number of CATI countries where issue…			Number of CAPI countries where issue…		
		…occurred	…is positively associated	…is negatively associated	…occurred	…is positively associated	…is negatively associated
Paradata	Interviewer workload	3	1	0	4	2	0
	Overlapping interviews	9	3	0	6	0	0
	Interviews in multiple PSUs	0	0	0	6	1	0
	Interviews between 10 pm and 7 am	3	0	0	4	1	0
	Short interviews	10	9	0	7	7	0
	Long interviews	10	0	0	7	0	0
Substantive results	No item missings	10	1	0	7	4	0
	High item missings	10	0	0	7	1	0
	Straightlining	10	1	0	7	1	0
	High match	10	0	0	7	2	0

Associations with QC indicators – short timestamps by country

Country		Paradata			Substantive data		
		Occurred	Positive association	Negative association	Occurred	Positive association	Negative association
CATI	N	2	1	0	4	0	0
	O	5	1	0	4	0	0
	F	4	3	0	4	0	0
	M	3	1	0	4	0	0
	E	3	2	0	4	1	0
	J	3	1	0	4	1	0
	D	3	1	0	4	0	0
	K	5	1	0	4	0	0
	H	4	1	0	4	0	0
	L	3	1	0	4	0	0
CAPI	P	5	3	0	4	2	0
	B	5	1	0	4	2	0
	G	4	1	0	4	0	0
	C	6	2	0	4	1	0
	Q	4	2	0	4	0	0
	A	4	1	0	4	2	0
	I	6	1	0	4	1	0

Associations with other QC indicators – long timestamps

		Number of CATI countries where issue…			Number of CAPI countries where issue…		
		…occurred	…is positively associated	…is negatively associated	…occurred	…is positively associated	…is negatively associated
Paradata	Interviewer workload	3	0	2	4	0	2
	Overlapping interviews	9	7	0	6	3	0
	Interviews in multiple PSUs	0	0	0	6	1	1
	Interviews between 10 pm and 7 am	3	0	1	4	1	0
	Short interviews	10	0	2	7	0	2
	Long interviews	10	4	0	7	4	0
Substantive results	No item missings	10	0	9	7	3	2
	High item missings	10	7	0	7	3	0
	Straightlining	10	0	0	7	0	0
	High match	10	0	0	7	2	0

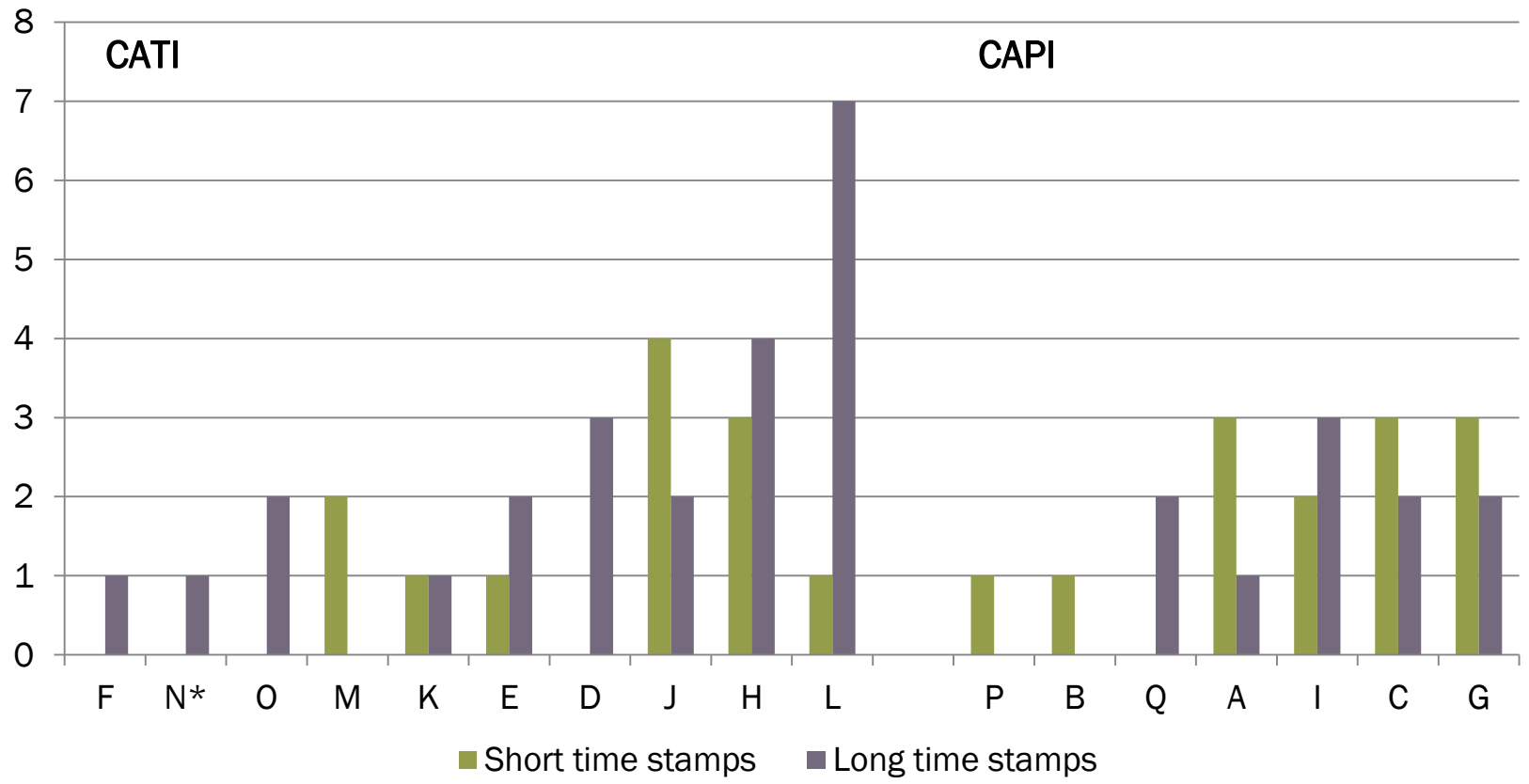
Associations with QC indicators – long timestamps by country

Country		Paradata			Substantive data		
		Occurred	Positive association	Negative association	Occurred	Positive association	Negative association
CATI	N	2	0	0	4	1	1
	O	5	2	1	4	0	0
	F	4	1	0	4	1	1
	M	3	1	0	4	1	1
	E	3	2	1	4	1	1
	J	3	0	1	4	0	1
	D	3	1	0	4	1	1
	K	5	2	0	4	1	1
	H	4	0	0	4	1	1
	L	3	2	0	4	0	1
CAPI	P	5	2	0	4	1	0
	B	5	1	0	4	2	1
	G	4	1	1	4	0	0
	C	6	1	1	4	1	0
	Q	4	1	2	4	1	1
	A	4	2	0	4	2	0
	I	6	1	1	4	1	0

Background characteristics of cases flagged based on timestamps

		Short timestamps			Long timestamps		
	Country	Gender	Age	Education	Gender	Age	Education
CATI	F					Older	ISCED 1 ↑
	O			ISCED 2 and 3 ↑			
	H		Older			Older	ISCED 1 ↑
	E					Older	
	L				Female ↑	Older	ISCED 1 ↑
	J					Older	ISCED 5 ↓
	D					Older	
	M		Older				
	N			ISCED 2 ↑		Older	ISCED 5 ↓
	K		Younger			Older	
CAPI	I					Younger	
	C				Male ↑	Older	
	P				Female ↑		
	G					Older	
	Q					Older	ISCED 1 ↑
	A		Younger				
	B						

Significant associations with substantive variables (N=14)



Controlling for gender, age and education

* For country N only 12 variables were available

Key findings (1)

- Short timestamps
  - Positively associated with short interviews in almost all countries, but they are not negatively associated with long interviews.
  - Not consistently associated with most of the other quality indicators.
    - In more than half of the CAPI countries short timestamps were associated with the absence of item non-response.
  - Respondents flagged by short timestamps tend not to differ from the general population

- Long timestamps
  - Positively associated with long interviews in about half of the countries, and negatively associated with short interviews in some countries.
  - Relatively frequently associated with high levels of item non-response and with overlapping interviews.
    - Associations found more often in CATI countries than in CAPI countries
  - Respondents flagged by long timestamps are older and less educated in many countries

Key findings (2)

- CAPI differs from CATI
  - Variability in item duration greater in CAPI countries than in CATI countries
    - More cases flagged in CAPI than in CATI
  - Different patterns in associations between timestamps and
    - other quality indicators
    - substantive questions

- Considerable country differences
  - In the extent to which timestamps are associated with other quality indicators.
    - In most countries the only associations found are with interview duration;
    - In 6 countries timestamp flags were associated with half or more of the other quality indicators
  - In the extent to which timestamps are associated with substantive questions.
    - In most countries no or very few associations were found
    - In some countries short timestamps were found to be associated with 3 or 4 of the 14 substantive variables tested.
    - In one country long timestamps were associated with 7 out of the 14 items tested.

Discussion

- **Issues with timestamps can indicate multiple things**
  - Short timestamps might indicate that an interview did not take place or a question block was not asked
  - Both short and long timestamps might indicate that an interviewer moved back an forth through the questionnaire
  - Long timestamps might indicate that an interview was interrupted and resumed later

- **Timestamp flags indicate different issues in different countries.**
  - In some countries they appear to indicate technical issues, that are not associated with any of the other quality indicators or with substantive variables
  - In other countries they appear to indicate interviewer behavior that is associated with other quality indicators and with substantive variables

- **Timestamps appear not to be very well suited for use as single indicators**
  - Need to be used in combination with other indicators, which can determine the severity of the quality concerns raised by timestamp flags

Looking ahead

- Timestamps are now a requirement for all modes and countries
  - At least on the section level
  - Allow us to compare the functionality of timestamps across regions and cultures
- Question level timestamps in addition to sectional time measures
  - Allows for assessment of question complexity and comprehension
- Distribution vs. hard cut-offs
  - Further develop procedures and metrics to determine outliers in terms of interview, section or item duration
- Speed measures – seconds per item
  - Further develop ways to use the average time spent on items more effectively in the assessment of data as well as the automated checking of interviewers work
- Working with local fieldhouses to better understand time measures
  - Strike right balance between local best practice and harmonization
  - Further raise awareness of importance of collecting high quality paradata

PewResearchCenter

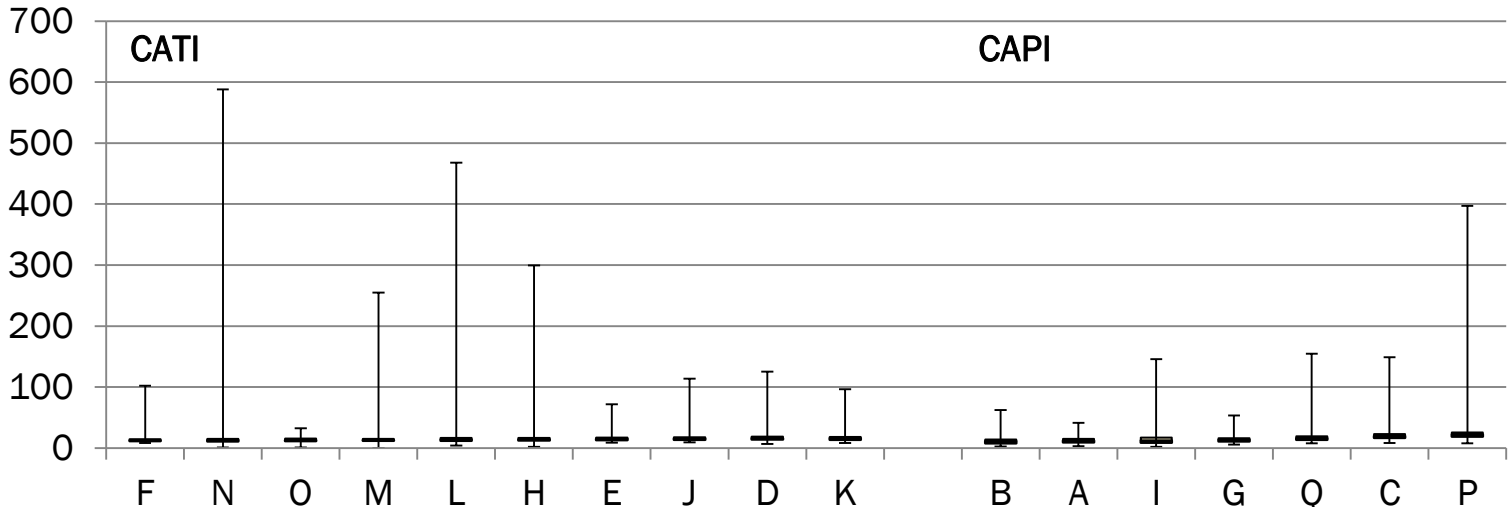
Questions or suggestions?

Gijs van Houten
Research Manager, Eurofound
Gijs.vanHouten@eurofound.europa.eu

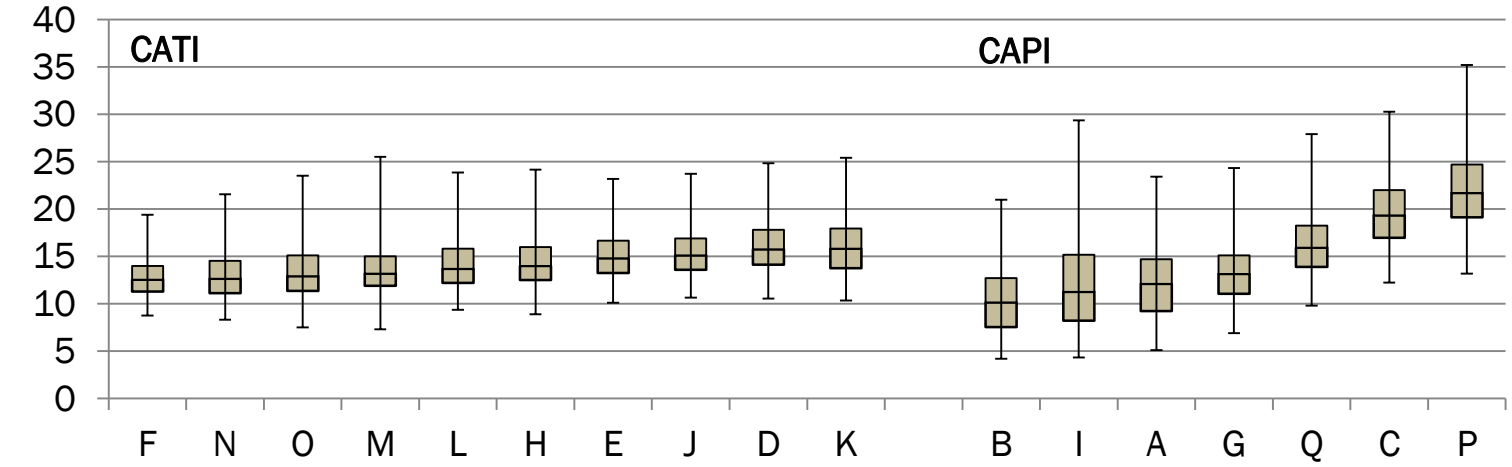
Steve Schwarzer
Research Methodologist, Pew Research Center
SSchwarzer@PewResearch.org

Distribution of average item duration (seconds)

All cases



5th-95th pctile



Using natural logs to determine relative extremes